CLAIMS

What is claimed is:

1. A system for identifying genes, comprising:

a pattern database comprising patterns of amino acids;

an input device for inputting a DNA sequence; and

a processor for processing said DNA sequence and said patterns to identify a putative gene.

2. The system according to claim 1, wherein said processor determines an open reading frame (ORF) in said DNA sequence.

3. The system according to claim 2, wherein said processor generates an amino acid translation for said ORF.

4. The system according to claim 3, wherein said processor identifies a match of a pattern from said pattern database in said amino acid translation.

5. The system according to claim 4, wherein said patterns are derived from a parent database comprising at least one amino acid sequence.

6. The system according to claim 4, wherein said patterns are derived from a parent database comprising at least one amino acid sequence fragment.

7. The system according to claim 4, wherein said patterns are derived by using a pattern discovery algorithm.

8. The system according to claim 4, wherein said patterns are derived by using the Teiresias algorithm.

9. The system according to claim 4, wherein said ORF comprises a portion of said DNA sequence between a start codon and a stop codon.

10. The system according to claim 4, wherein said ORF is reported as a putative gene when a predetermined number of pattern matches is identified in said amino acid translation.

11. The system according to claim 4, wherein each pattern is assigned a weight depending upon a relevance of said pattern in determining whether said ORF comprises a putative gene.

12. The system according to claim 4, wherein said ORF is reported as a putative gene when the sum of weights corresponding to all patterns with matches in said amino acid translation exceeds a predetermined threshold.

13. The system according to claim 4, wherein said match is identified using a predetermined pattern matching algorithm.

14. The system according to claim 4, further comprising:

a memory device for storing data and instructions to be executed by said processor.

15. The system according to claim 4, further comprising:

a display device for displaying an output from said processor.

16. A method of identifying genes, comprising:

providing a pattern database comprising patterns of amino acids;

determining an open reading frame (ORF) in a DNA sequence;

generating an amino acid translation for said ORF; and

identifying a match of a pattern in said amino acid translation.

17. The method according to claim 16, wherein said pattern database is generated from a database comprising at least one amino acid sequence.

18. The method according to claim 16, wherein said pattern database is generated from a database comprising at least one amino acid sequence fragment.

19. The method according to claim 16, further comprising:

identifying said ORF as a putative gene when a predetermined number of pattern matches is identified in said amino acid translation.

20. The method according to claim 16, further comprising:

assigning a weight to each pattern depending upon a relevance of said pattern in determining whether said ORF is a putative gene.

21. The method according to claim 16, further comprising:

displaying said match of said pattern in said amino acid translation.

22. The method according to claim 16, wherein said pattern database is generated using the Teiresias algorithm to derive said patterns from a parent database.

23. A programmable storage medium tangibly embodying a program of machine-readable instructions executable by a digital processing apparatus to perform a method for identifying genes, said method comprising:

providing a pattern database comprising patterns of amino acids;

determining an open reading frames (ORF) in a given DNA sequence;

generating an amino acid translation for each ORF; and

identifying a match of a pattern in said amino acid translation.